Incremental Difference as Feature for Lipreading

Pravin L Yannawar[1], Ganesh R Manza[2], Bharti W Gawali[3], Suresh C Mehrotra[4]

[1,2,3,4] Department of Computer Science and Information Technology,
Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India
pravinyannawar@gmail.com, ganesh.maza@gmail.com, drbhartirokade@gmail.com, scmehrotra@yahoo.com

Abstract— **This paper represents a method of computing incremental difference features on the basis of scan line projection and scan converting lines for the lipreading problem on a set of isolated word utterances. These features are affine invariants and found to be effective in identification of similarity between utterances by the speaker in spatial domain.**

Keywords- **Incremental Difference Feature, Euclidean distance, Lipreading.**

I. Introduction

An Automatic speech recognition (ASR) for well defined applications like dictations and medium vocabulary transaction processing tasks in relatively controlled environments has been designed. It is observed by the researchers that the ASR performance is far from human performance in variety of tasks and conditions, indeed ASR to date is very sensitive to variations in the environmental channel (non-stationary noise sources such as speech babbled, reverberation in closed spaces such as car, multi-speaker environments) and style of speech (such as whispered etc)[1].

Lipreading is an auditory, imagery system as a source of speech and image information. It provides the redundancy with the acoustic speech signal but is less variable than acoustic signals; the acoustic signal depends on lip, teeth, and tongue position to the extent that significant phonetic information is obtainable using lip movement recognition alone [2][3]. The intimate relation between the audio and imagery sensor domains in human recognition can be demonstrated with McGurk Effect [4][5]; where the perceiver "hears" something other than what was said acoustically due to the influence of conflicting visual stimulus. The current speech recognition technology may perform adequately in the absence of acoustic noise for moderate size vocabularies; but even in the presence of moderate noise it fails except for very small vocabularies[6][7][8][9]. Humans have difficulty distinguishing between some consonants when acoustic signal is degraded.

However, to date all automatic speech reading studies have been limited to very small vocabulary tasks and in most of cases to very small number of speakers. In addition the numbers of diverse algorithms have been suggested in the literature for automatic speechreading and are very difficult to compare, as they are hardly ever tested on any common audio visual databases. Furthermore, most of such databases are very small duration thus placing doubts about generalization of reported results to large population and tasks. There is no specific answer to this but researchers are concentrating more on speaker independent audio-visual large vocabulary continuous speech recognition systems [10].

Many methods have been proposed by researcher's in-order to enhance speech recognition system by synchronization of visual information with the speech as improvement on automatic lipreading system which incorporates dynamic time warping, and vector quantization method applied on alphabets, digits. The recognition was restricted to isolated utterances and was speaker dependent [2]. Later *Christoph Bregler (1993)* had worked on how recognition performance in automated speech perception can be significantly improved & introduced an extension to existing Multi-State Time Delayed Neural Network architecture for handling both the modalities that is acoustics and visual sensor input [11]. Similar work has been done by Yuhas *et.al* (1993) & focused on neural network for vowel recognition and worked on static images [12].

Paul Duchnowski et.al (1995) worked on movement invariant automatic lipreading and speech recognition [13], Juergen Luettin (1996) used active shape model and hidden markov model for visual speech recognition [14]. *K.L.* Sum *et.al* (2001) proposed a new optimization procedure for extracting the point-based lip contour using active shape model [16]. Capiler (2001) used Active shape model and Kalman filtering in spatiotemporal for noting visual deformations [17]. Ian Matthews *et.al* (2002) has proposed method for extraction of visual features of lipreading for audio-visual speech recognition [18]. Xiaopeng Hong *et.al* (2006) used PCA based DCT features Extraction method for lipreading [19]. Takeshi Saitoh *et.al* (2008) has analyzed efficient lipreading method for various languages where they focused on limited set of words from English, Japanese, Nepalese, Chinese, Mongolian. The words in English and their translated words in above listed languages were considered for the experiment [20]; Meng Li *et.al* (2008) has proposed a Novel Motion Based Lip Feature Extraction for Lipreading problems [21].

The paper is organized in four sections. Section I deals with introduction and literature review. Section II deals with methodology adopted. Section III discusses results obtained by applying methodology and section IV contains conclusion of the paper.

DOI: 01.IJIT.01.01.66

II. Methodology

The System takes the input in the form of video (moving picture) which is comprised of visual and audio data as shown in Figure 1. This will act as an input to the audio visual speech recognition. The samples from the subjects having devnagari language as mother tongue have been collected. The isolated words of city names in have been pronounced by the speakers.

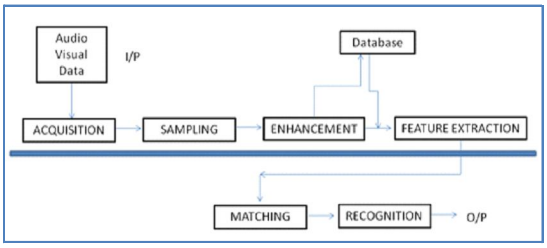


Figure 1: Proposed Model

The samples from female subject have been chosen. Each speaker or subject is requested to begin and end each letter utterances for isolated city names with their mouth in closed-open-close position. No head movement is allowed and speakers have been provided with close up view of their mouth and urged to do not move face out of the frame. With these constraints the dataset is prepared. This video input was acquired by acquisition phase and passed to sampler which samples video into frames. The video samples of subject were sampled by sampler. This sampling of frame was done with the standard rate of 32 frames per second. Normally the Video input of 2 seconds was recorded for each subject. When these samples were provided to sampler; it has produced 64 images for utterance and was considered as image vector 'I' of size 64 images and shown in Figure 2.

The image vector 'I' has to be enhanced because images in vector 'I' are dependent on lighting conditions, head positions etc. The registration or realignment of image vector 'I' was not necessary. The entire sample collected from subject was in constrained environment, as discussed above.

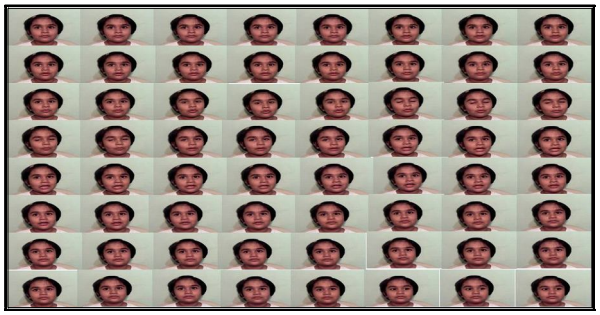


Figure 2:Subject with utterance of word "MUMBAI", Time 0.02 Sec @ 32Fps

Image vector 'I' was processed for color to gray and further to binary, with histogram equalization, background estimation and image morphological operation by defining structural element for open, close, adjust operations. The outcome of this preprocessing is shown in Figure 2.

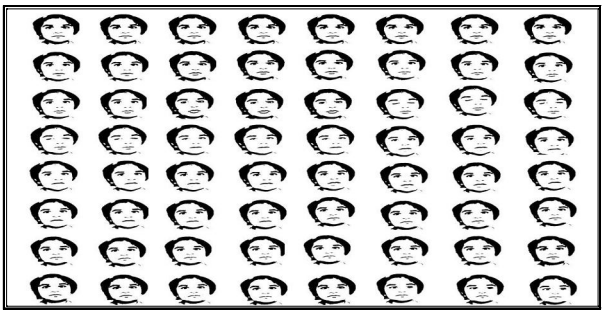


Figure 3: Subject with utterance of word "MUMBAI", Time 0.02 Sec @ 32Fps, Gray to Binary image conversion using Morphological Operation with Structure Element 'Disk'

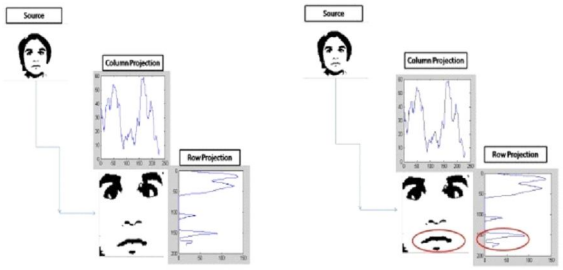
A. Region of Interst:

The identification of Region of Interest (ROI) from binary image the scan line projections of row as R(x), columns as C(y); were computed as a vectors with respective to every frame. The image from vector is represented by two dimensional light intensity function F(x,y) returning amplitude at an coordinate x,y

$$R(x) = \sum_{x=1}^{m} \sum_{y=1}^{n} F(x, y) \qquad \text{.....(1)}$$

$$C(y) = \sum_{y=1}^{n} \sum_{x=1}^{m} F(y, x) \qquad \text{.....(2)}$$

This process suggests the area for segmentation of eyes, nose & mouth from the every image of vector. This was found to be helpful in classifying open-close-open mouth of the subject as well as some geometrical features such as height, width of mouth in every frame can easily be computed.



(a) (b)

Figure 4 (a) Horizontal (row) and vertical (column) scan line Projections of Face to locate facial components like eyes, eyebrows, nostrils, nose, mouth, (b) Isolation of Mouth Region

The masking was done so as to reduce workspace. When the R(x) and C(y) were plotted, the plot represents the face components like eyes, nose, and mouth. The masking containing mouth region was framed in accordance with very first image in vector 'I', this was accomplished by computing horizontal scan line projection (row projections) and vertical scan line projections (column projections) as discussed above. On source vector 'I', it was observed that the face components like eyes, eyebrows, nose, nostrils, and mouth could be easily be isolated. The region of interest, that was mouth can easily be located as show in Figure 4 (a) and it was very easy to segment into three parts like eyes, nose, mouth, as show in Figure 4 (b). The masking remained constant for all remaining images of the vector and window coordinate containing mouth was fixed for Mask. This

DOI: 01.IJIT.01.01.66

was applied to all frames of Image vector so that mouth frame from source image was extracted. The result of windowing operation was resulted in vector called 'W' as shown in Figure 5

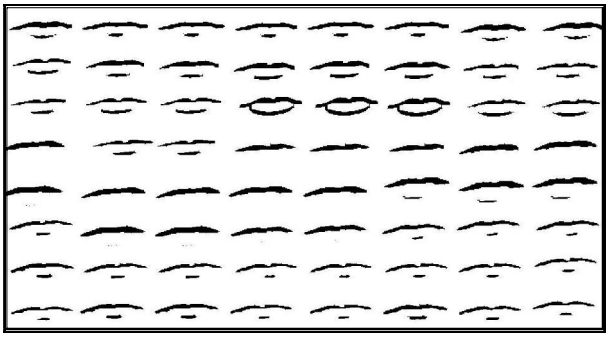


Figure 5: Masking result of Subject with word "MUMBAI", Time 0.02 Sec @ 32Fps

Feature extraction

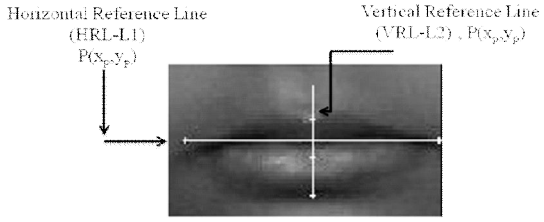


Figure 6: HRL and VRL for identification for Incremental Difference

To extract the lip features from frame vector there have been two approaches. The First one is low level analysis of image sequence data which does not attempt to incorporate much prior knowledge. Another approach is a high level approach that imposes a model on the data using prior knowledge. Typically high level analysis uses lip tracking to extract lip shape information alone. The feature extraction was carried out by low level analysis by directly processing image pixels and is implicitly able to retrieve additional features that may be difficult to track such as teeth and tongue [22].

Low level analysis is adopted in order to compute the features. The Horizontal Reference Line (HRL) and Vertical Reference Line (VRL) for the lip are plotted. The points for HRL and VRL have been chosen from scan line projection vectors that is R(x) and C(y). The initial values for P_1 which were the midpoint for HRL was calculated as

$x_p = (x_2 - x_1)/2$ and $y_p = (y_2 - y_1)/2$ where x_2, x_1 are the co-ordinates obtained from R(x) and y_2,y_1 are obtained for C(y). The initial values for P_2 which is the midpoint for VRL is calculated as that of P_1. Therefore to obtain exact midpoint of HRL with reference to P_1 at (x_p, y_p) and VRL with reference to P_2 at (x_p, y_p) ,the HRL & VRL are represented by the line implicit function with coefficient a, b and $c : F(x, y) = ax + by + c = 0$ (the b coefficient of y is unrelated to the y intercept B in the slope intercept form). If $dy = y_2 - y_1$ and $dx = x_2 - x_1$ the slope intercept form can be written as

$$y = \frac{dy}{dx}x + B$$

Therefore

$$F(x, y) = dy \cdot x - dx \cdot y + B \cdot dx = 0$$

Here $a = dy, b = -dx$ and $c = B \cdot dx$ in the implicit form as it is important for the proper functioning of the midpoint of HRL to choose 'a' to be positive; so that it meets this criterion if dy is positive, since $y_2 > y_1$.

To calculate midpoint criterion for HRL and VRL as for the pixel point P_1 and P_2 as we need to compute $F_{HRL}(M)$ and $F_{VRL}(M)$ as

$$F_{HRL}(M) = F(x_p + 1, y_p + \frac{1}{2}) \quad \text{.....(3)}$$

$$F_{VRL}(M) = F(x_p + 1, y_p + \frac{1}{2}) \quad \text{.....(4)}$$

The decision is based on the value of the function at $(x_p + 1, y_p + \frac{1}{2})$ it is necessary to define decision variable for HRL and VRL respectively

$$d = F(x_p + 1, y_p + \frac{1}{2}) \quad \text{.....(5)}$$

Therefore by definition

$$d = a(x_p + 1) + b(y_p + \frac{1}{2}) + c \quad \text{.....(6)}$$

Conditions

If d>0 then we choose pixel NE (North East)
If d<0 then we choose pixel E (East)
If d=0 then we can choose either, recommended to choose E

The location of M is on whether we chose E or NE, if E is chosen, and then M is incremented by one step in x direction then

$$d_{new} = F(x_p + 2, y_p + \frac{1}{2}) = a(x_p + 2) + b(y_p + \frac{1}{2}) + c \quad \text{....(7)}$$

But,

$$d_{old} = a(x_p + 1) + b(y_p + \frac{1}{2}) + c \quad \text{.....(8)}$$

If NE is chosen, M is incremented by one step each in both x and y direction then

$$d_{new} = F(x_p + 2, y_p + \frac{3}{2}) = a(x_p + 2) + b(y_p + \frac{3}{2}) + c \quad \text{...(9)}$$

By equation (3) and (4) [23] with support of decision variable new coordinates for P_1(x, y) for HRL and P_2(x,y) for VRL are computed. The pixel P1 was at the middle of HRL and pixel P_2 lies at the middle of VRL. The difference between the P_1 and P_2 is considered as incremental difference feature and will be unique feature for the frame. This feature is invariant to scale, rotation and scaling. This difference is computed for all frames for utterance of word and stored in vector; this vector will be referred as feature vector for the word. The feature vector will contain the information of all samples of word such as {AURANGABAD,MUMBAI, PARBHANI, KOLHAPUR, and OSMANABAD}.

III. Result And Discussion

The midpoint (M) for HRL and VRL has been chosen on the basis of above discussed method. The pixel P_1 and P_2 are with new (x_p, y_p) respectively and are marked as landmark

DOI: 01.IJIT.01.01.66

points on the every frame of vector as shown in Figure 7.

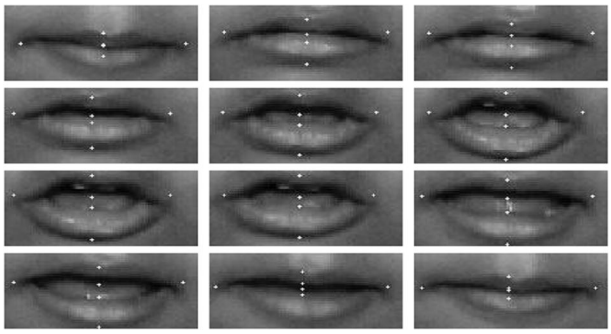


Figure 7: Marking of all landmark points

The pixel difference between P_1 and P_2 was recorded as feature of the frame and similar difference with respect to all frames of vector have been computed and stored in the feature vector. The feature vector corresponding to all utterance of the word 'AURANGABAD' is formed, their mean feature vector is also calculated. The Euclidean distance between mean feature vector and computed feature vector has been computed and represented in Table I. From table I, it is observed that the sample 1 and sample 2 of the word 'Aurangabad' are found to be similar and sample 4 and sample 5 are also found to be similar, the sample 8 and sample 9 are same. The similar kinds of results were obtained for the other samples of the words uttered by the speaker. The Table II represents the Euclidean distance metrics for the word 'AURANGABAD' by the speaker1. The Graph I shows similarity between the Maxima and Minima from the feature vectors of Sample 1 and Sample 2 of the word 'AURANGBAD' uttered by the speaker 1 and the Graph II shows how two words that is 'AURANGABAD' and 'KOLHAPUR' uttered by same speaker are different on the basis of computed feature vectors and Maxima and Minima observed in graph II. The mean feature vectors of these words are plotted. This feature vector is formed with the help of incremental difference procedure

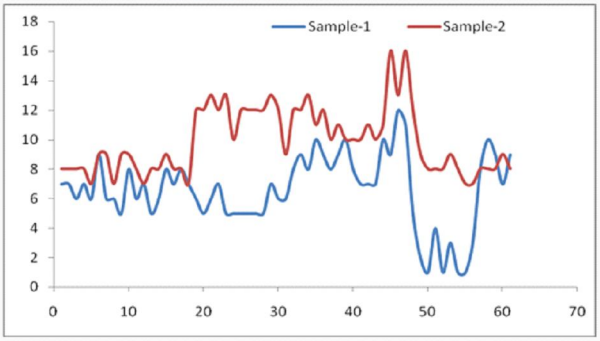
TABLE I
EUCLIDEAN DISTANCE BETWEEN MEAN FEATURE OF EACH WORD AND FEATURE VECTOR FOR THE WORD UTTERED BY SPEAKER 1

Sample	Aurangabad	Mumbai	Kolhapur	Parbhani
1	25.74	16.34	22.49	49.40
2	25.86	23.22	61.12	24.02
3	43.89	34.58	56.89	19.26
4	31.87	46.66	62.15	22.89
5	31.17	28.23	53.30	24.06
6	33.61	13.60	26.70	37.47
7	30.18	25.24	40.29	47.30
8	27.67	25.34	33.11	39.03
9	27.67	23.41	41.23	38.25
10	47.39	0.00	45.45	45.16

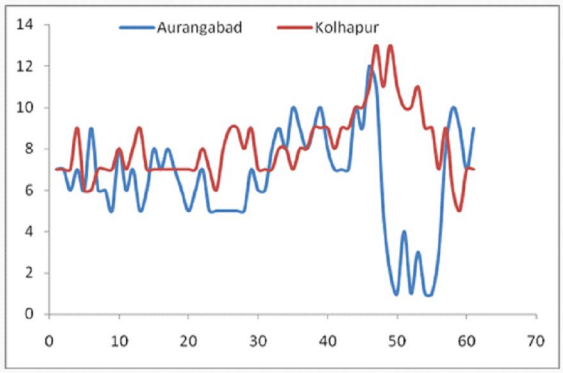
TABLE II
EUCLIDEAN DISTANCE MATRIX FOR WORD AURANGABAD

Word	1	2	3	4	5
1	0	25.74	25.86	43.89	31.87
2	25.74	0	31.17	33.61	30.18
3	25.86	31.17	0	27.67	27.67
4	43.89	33.61	27.67	0	47.39
5	31.87	30.18	27.67	47.39	0

GRAPH I. SIMILARITY BETWEEN THE UTTERENCES OF WORD 'AURANGABAD' BY THE SPEAKER 1



GRAPH II. DIFFERENCE BETWEEN UTTERENCES OF WORD 'AURANGABAD' AND 'KOLHAPUR' BY SPEAKER 1



The similar results are observed from the other samples of speakers.

IV. CONCLUSION

The incremental difference a novel method for feature extraction for audio-visual speech recognition and it is found to be suitable for the enhancement of speech recognition. This method helps in differentiating the words spoken by the speaker.

ACKNOWLEDGEMENT

The author's would like to thank of the university authorities for providing all the infrastructures required for the experiments.

REFERENCES

[1] J R Deller, Jr. J G Proakis and J.H L Hansen, Discrete-Time Processing of Speech Signals, Macmillan Publishing Company, Englewood cliffs, 1993.

[2] Eric Petjan, Bradford Bischoff, and David Bodoff, An Improved automatic lipreading system to enhance speech recognition, Technical Report TM 11251-871012-11, AT&T Bell Labs, Oct. 1987

DOI: 01.IJIT.01.01.66

[3] Finn K.I, An investigation of Visible lip information to be used in Automated speech recognition, Ph.D Thesis, George-Town university, 1986

[4] Macdonald J and H MacGurk, Visual influences on speech perception process, Perception and Psychophysics, Vol 24 pp 253-257, 1978

[5] MacGurk H and Macdonald J, Hearing lips and seeing voices, Nature vol 264, pp 746-748, Dec 1976.

[6] Paul D.B, Lippmann R.P, Chen Y and Weinstein C.J, Robust HMM based technique for recognition of speech produced under stress and in noise, Proceeding Speech Tech. 87, pp 275-280, 1987

[7] Malkin F.J, The effect on computer recognition of speech when speaking through protective masks, Proceeding Speech Tech. 87, pp 265-268, 1986

[8] Meisel W.S, A Natural Speech recognition system, Proceeding Speech Tech. 1987, pp 10-13,1987

[9] Moody T, Joost M and Rodman R, A Comparative Evaluation a speech recognizers, Proceeding Speech Tech. 87, pp 275-280, 1987.

[10] Chalapathy Neti, Gerasimos Potaminos, Juergen Luettin, Ian Matthews, Herve Glotin, Dimitra Bergyri, June Sison, Azad Mashari and Jie Zhou, Audio-Visual Speech Recognition, Workshop 2000 Final report, Oct 2000.

[11] Christopher Bergler, Improving connected letter recognition by lipreading, IEEE, 1993

[12] B P Yuhas, M H Goldstien and T.J Sejnowski, Integration of acoustic and visual speech signals using neural networks, IEEE Communication Magazine.

[13] Paul Duchnowski, Toward movement invariant automatic lipreading and speech recognition, IEEE, 1995

[14] Juergen Luetin, Visual Speech recognition using Active Shape Model and Hidden Markov Model, IEEE,1996

[15] A. K Jain, Statistical Pattern Recognition: A Review, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 22, No 1, January 2000.

[16] K.L Sum, W H Lau, S H Leung, Alan W. C. Liew and K W Tse, A New Optimization procedure for extracting the point based lip contour using Active Shape Model, IEEE International confernece on Acoustics, Speech and Signal Processing, Salt Lake City, UT, USA, pp 1485-1488, 7th-11th May 2001

[17] A Capiler, Lip detection and tracking, 11th International Conference on Image Analysis and Processing (ICIAP 2001).

[18] Ian Matthews, T F Cootes, J A Banbham, S Cox, Richard Harvey, Extraction of Visual features of Lipreading, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol 24, No 2, February 2002.

[19] Xiaopeng Hong, Hongxun Yao; Yuqi Wan; Rong Chen, A PCA based Visual DCT feature extraction method for lipreading, International conference on Intelligent Information hiding and multimedia signal Processing, 2006.

[20] Takeshi Saitoh, Kazutoshi Morishita and Ryosuke Konishi, Analysis of efficient lipreading method for various languages, International Conference on Pattern Recognition (ICPR) , Tampa, FL, pp 1-4, 8th-11th Dec 2008

[21] Meng Li, Yiu-ming Cheung, A Novel motion based Lip Feature Extraction for lipreading, IEEE International conference on Computational Intelligence and Security, pg.no 361-365, 2008.

[22] Ian Mathews, Features for Audio Visual Speech Recognition, Ph.D Thesis, School of Information Systems, University of East Anglia, 1998

[23] James D Foley, Andries Van Dam, Steven K Fiener, John F Huges, Computer Graphics Principal and Practice,Pearson Education Asia, Second Edition ISBN 81-7808-038-9

DOI: 01.IJIT.01.01.66